DOCUMENT RESUME

ED 115 078 FL 007 035

AUTHOR Beardsmore, H. Baetens
TITLE Testing Oral Fluency. Rapport d'Activites de l'Institut de Phonetique, 1971-1972 (Report of the Activities of the Institute of Phonetics).
INSTITUTION Universite Libre de Bruxelles (Belgium). Institut de Phonetique.
PUB DATE Oct 72
NOTE 14p.; Paper presented at the Annual Meeting of l'Association Belge de Linguistique Appliquee (September 25, 1972)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS English (Second Language); *Language Fluency; *Language Instruction; Language Skills; *Language Tests; Oral Expression; Scoring; Second Language Learning; *Speech Skills; *Testing
IDENTIFIERS *Communicative Competence

ABSTRACT

A description is given of experiments involving the standardization of aspects of oral fluency testing. Oral fluency is understood to imply a "communicative competence" requiring an ability to formulate accurate and appropriate utterances of more than one sentence in length. Throughout the test emphasis is laid on ease of application and relevance. In the quest for greater objectivity in testing procedures, linguistic, methodological and technical aspects are considered, together with tentative criteria for reducing the discrepancy ratio between scorers. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED115078

TESTING ORAL FLUENCY*

Summary

A description is given of experiments involving the standardisation of different aspects of oral fluency testing. Oral fluency is understood to imply a "communicative competence" requiring an ability to formulate accurate and appropriate utterances of more than one sentence in length. Throughout the test emphasis is laid on ease of application and relevance. In the quest for greater objectivity in testing procedures, linguistic, methodological and technical aspects are considered, together with tentative criteria for reducing the discrepancy ratio between scorers.

Résumé

Quelques expériences sur la standardisation de plusieurs aspects de l'appréciation de la fluidité verbale sont décrites. Par fluidité verbale on entend une "compétence communicative" nécessitant la formulation d'expressions correctes et appropriées d'une longueur de plus d'une phrase. La facilité de l'application du test et sa pertinence sont soulignées. Dans la recherche pour une plus grande objectivité dans les procédures de l'examen, les aspects linguistiques, méthodologiques et techniques sont pris en considération, de même qu'une tentative pour réduire le taux de divergence entre les différents correcteurs.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Institut de Phonétique
Université Libre de Bruxelles
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER"

FL007035

* Texte d'un exposé présenté le 25.IX.1972 à la réunion annuelle de l'Association Belge de Linguistique Appliquée (ABLA).

Introduction

The quest for methods of testing oral production in an objective standardised way has been going on with greater or lesser success for several years now (cf. Dodson, 1963; Pimsleur, 1966; Grayshon, 1968). However, much of the material published on this difficult aspect of foreign language testing has tended to reveal more about the shortcomings of standardised techniques when compared with the interview situation or else shows signs of methodological, terminological or functional inadequacies.

Quite often these shortcomings have been caused by an insufficient awareness of the relationship between the test situation and the teaching programme which it is supposed to reflect. In many cases the goals of a teaching programme are represented as encouraging the active, meaningful, socially appropriate, creative use of language while the test battery in use measures discrete points of the learner's foreign language knowledge in the hope that success in the latter faithfully reflects ability at the former (1). Where an awareness exists of the danger in equating performance at discrete-point tests with actual performance in a real life situation the interview is generally considered the best means of judging proficiency. And yet dissatisfaction with traditional interview techniques is widespread as is the feeling of frustration at the inability to replace interviews by anything which adequately meets criteria of reliability and validity.

By combining aspects of both types of testing, i.e. the global approach and the minute approach, into a test battery it should be possible to combine the advantages of the interview situation with the standardisation and objectivity of the discrete-point test without at the same time running the risk of inadequacy or irrelevance.

This paper hopes to show what hesitant, if modest, progress is being achieved in this direction, and more specifically, in the standardising of tests of oral production. What is presented is not a fool-proof technique applicable in a variety of testing situations, but an indication of some of the possibilities available, together with their concomitant limitations.

---

1) [illegible]bovits (1971, pp. 78-79).

Work on the standardisation of tests of oral production grew from an experiment carried out at the Université Libre de Bruxelles in which recorded interviews were scored by different and by the same examiners at different periods of time. Not only were there discrepancies in the scoring when the recordings were judged by different examiners (as might be expected), but the same examiners also scored the same candidates differently after a three week lapse of time between the original interview and the second hearing. As a result of this experience it was decided to produce new tests of oral production which, in the initial stages, would be complementary to interview testing in the hope of eventually replacing the latter by standardised tests applicable to large populations in a manner that was not time consuming.

The starting point for any work on the eventual replacement of interviews had of necessity to be an analysis of what factors came into play in the test situation. These were then critically examined and classified into those which could be readily quantified and standardised (and were therefore more likely to be treated objectively) and those which could not (and were therefore open to subjectivity). A typological inventory of language specimens and their linguistic components led to an awareness of the restrictions these imposed on the test situation. A battery of tests was designed to test discrete points while the information sought for in the interview was analysed and broken down into two sections.

Quantifiable elements of the interview situation (e.g., the use of basic structures involving the correct, spontaneous manipulation, in a rigorously defined natural context of such elements as tense markers, pronoun changes, word order, etc.) have been described in "A Test of Spoken English" (2). This test, which has successfully been applied to more than a thousand candidates over the past three years, was based on ideas that rejected the behaviourist principle of taking into account only measurable linguistic expression where this quantifiable mechanistic definition led to the atomisation and dehumanisation of language (3). The test was

---

2) H. Baetens Beardsmore & A. Renkin (1971).

3) Cf. Jakobovits' claim in this respect that scientific exactitude on these mechanistic lines is inversely proportional to realistic concerns. (Jakobovits, 1971, p. 189).

based on a global approach offering advantages of validity and relevance and therefore approximated a real-life language situation.

However the "Test of Spoken English", although successfully providing information as to a student's capacity to react spontaneously and accurately to natural sequences of speech, did not provide information about the student's ability to manipulate connected speech of more than one sentence in length. In order for the battery of spoken English tests to be complete it was felt necessary to test not only the student's linguistic competence, but also his communicative competence, by which we understand the ability to produce contextually appropriate utterances (4). This, it was felt was one of the important aspects of the interview situation, and an element of linguistic performance which, although more difficult to standardise, could not be neglected. Consequently it was decided to attempt to discover a means of testing oral fluency, where fluency represents not what certain psycholinguists would seem to consider, i.e. the speed of response (cf. Macnamara, 1969), but an ability to manipulate connected speech. More precisely :

> "Oral fluency requires the ready availability of this communicative competence for the formulation of appropriate utterances in real time, involving a strategy for the elaboration of sentence structures, as well as the selection and insertion of lexical items. Individual sentences must be integrated into connected discourse"(5).

---

4) Campbell & Wales (1970, pp. 246-247) distinguish between what they call "linguistic competence", made up of "competence$_1$" (capacity or ability in any sphere, reflecting underlying capacity), competence$_2$ (Chomskyian "competence as opposed to "performance"), and "competence$_3$", which is the ability to produce contextually appropriate utterances. This "competence$_3$" would seem to meet Jakobovits' (1971, p.156) definition of "communicative competence", which includes linguistic meaning, implicit meaning and implicative meaning. Spencer (1972) likewise emphasises the function of language within the speech community and the communicative competence in social interaction through language.

5) AILA Bulletin n° 1 (1972, p.9).

The Preliminary Project

Factors that had to be taken into consideration were administrative, technical and scoring factors, as well as those of standardising the linguistic content. Experiments were carried out on a large population of candidates where students were requested to speak on a given topic for a restricted amount of time after having received specific instructions and an opportunity to prepare themselves a few minutes beforehand.

In the preliminary project students were requested to persuade the listener, on tape, to do something as suggested by a series of topics at their disposal. The aim was to restrict topic, time available and type of context-determined structure likely to be used, as in an interview situation, while at the same time giving scope for variation in the selection of linguistic and factual components relevant to the presentation of ideas. In this way the creative aspect of constructing grammatical phrases was encouraged while still defining certain limits to serve as a framework of reference for the examiner. In this early stage scorers were given no instructions as to marking criteria other than to go about correcting in the same way as they would in a classical interview situation. This was justified by the fact that at this moment in the experiment it was felt that marking techniques were the greatest problem and could only be solved with some hope of success if other variables had been measured and taken into account previously.

Immediate inadequencies appeared in both methodological and technical aspects. Requesting students to persuade a listener to do something limited the variety of relevant topics that could be offered for selection. Moreover, structures were of necessity exceedingly rigid and limited to those determined by the argumentative nature of the topics, e.g. hypothetical conditionals. To overcome these inadequacies it was decided to request students to give the listener his opinions about a particular topic. Choice of themes was organised in such a way that a greater variety of structures might be elicited. Narrative topics were offered which might predispose the speaker to use past and present verb forms, present and future, present and conditional, etc. However, topics were at times found to be too general leaving students with very little to say and therefore rendering the test counter-productive. To allay this course-orientated topics were offered.

The timing of the test in the initial stages proved to be another perturbing factor since students were allowed too much time. The four minutes of blank tape on which they were to develop their ideas led to excessive repetition, to students running out of things to say, or else to panic-stricken efforts to fill in the time available leading to mistakes that more than counterbalanced the commendable efforts made initially.

In spite of all the teething troubles it was felt that sufficient measurable sustained speech was being produced to justify continuing the experiment.

The Present Test

a) Administration

The test of oral fluency at present in use has been taken by 396 students since September 1971. Administration takes place in the language laboratory with complete control of machines from the console. The student's only pre-occupation is therefore a linguistic one.

Particular attention is paid to the simplicity and clarity of instructions since it must always be borne in mind that the mechanics of the testing situation are such that students have no means of making up lost ground caused by an inadequate appreciation of what is required from them. Students are told that they should select one of three given topics, that they have two minutes in which to gather their ideas and make notes if desired, and that speaking-time will be limited to one minute forty-five seconds, which should be fully used. (This was found to be the optimum time for our particular population, at a given level. Trials with other populations showed that different time limits could be imposed at different levels of ability. An analysis of the interview situation had revealed that the time normally available for students to develop ideas in uninterrupted discourse (the objective in this test) as opposed to that taken up by direct questions and answers amounted to roughly two minutes).

Small blank cards are issued to students on which they write their names (for later indentification of tapes) and on which brief notes, to serve as memory aids, can be jotted. The size of the cards guarantees the brevity of the notes thereby preventing reading aloud of carefully prepared sentences.

The number of subjects is limited to three for reasons of standardisation and

cross-comparison. Topics are course-orientated but so worded that they do not allow for the straight regurgitation of language material learnt parrot-fashion from courses. The important feature of the test is its creative aspect (as in the interview situation) where students are required to re-apply structures and lexis acquired during the year. Topics are presented in a way that will hopefully predispose students to use particular tenses, structures, lexis, etc., and are equalised, in so far as this is possible, for difficulty of structure likely to be elicited.

b) Examples of Topics

1. Tell me about working and living conditions in North and South Belgium and how they have changed over the past fifty years
2. Tell me about the changing shape of transport in Belgium.
3. What advice would you give someone who wanted to invest a large amount of capital in Belgium?

c) Timing

The test takes ten minutes to administer to twenty-two students in the language laboratory. Scoring takes place at the examiners' convenience, each tape requiring approximately four minutes for correction.

d) Student Preparation

Any exam situation requires from examiners an acute awareness of the stress phenomena which may intervene to perturb student performance during examinations. Testing in language laboratories can aggravate the stress situation and if tests are to be at all valid students must be led to accept the machine-determined circumstances as natural and unimposing.

The limitation of time and topic would not normally worry candidates unduly in an interview situation, but an awareness of complete helplessness before a machine, coupled with the inability to go back and cover lost ground, can create hostility to the technique, if not the content matter of the examination. In order to minimise these problems students should be led to consider the examination as a simple extension of language-learning activities as practised throughout the course. It is absolutely essential that students receive training in the technique beforehand and this can be useful both as a teaching aid (e.g. for diagnostic testing) and

for accustoming students to the exam situation.

Throughout the year students are given trial tests where they are encouraged to speak spontaneously from summary notes, to vary their structures and lexis as much as possible within the restrictions imposed by the context and to talk onto their tape-recorders without hesitancy. The final examination should be conducted in the presence of and under instructions from the class teacher, so that the student feels the personal involvement of talking to a real human being.

e) Marking

The most difficult problem in testing oral fluency is of course that of scoring. Since the interview situation is the normal way in which the level of oral fluency is generally ascertained it was felt necessary to make an analysis, in so far as this was possible, of the factors applied in scoring during interviews. It was hoped that these factors would then make up the minimum criteria to be applied in the recorded test situation. Objectiveness and standardisation were not to be raised to the status of shibboleths capable of precluding any hopes of progress or success in the working out of the test. The attitude taken was that expressed in the Council of Europe's Symposium on "The Linguistic Content, Means of Evaluation and their Interaction in the Teaching and Learning of Modern Languages in Adult Education", where Gorosch suggests that the definition of the word objective should be established for humanistic purposes, for human global behaviour, of necessity implying some kind of subjective judgement(6). Nickel (7) is of a similar opinion, namely, that in the present state of our knowledge, objective testing can in many situations be no more than reduced subjective testing. At the moment the tool in use is blunt and imperfect but it does tend to measure what we want to measure, whereas many atomised language proficiency tests are extremely perfected and yet do not necessarily measure what is wanted (8).

---

6) Gorosch (1971, p. 2).

7) Nickel, G., in an unpublished communication read at the Institut de Phonétique de l'Université Libre de Bruxelles on "Error Analysis", 10/5/1972.

8) Gorosch (1971, p. 5).

In order to obtain some degree of standardisation scoring criteria were carefully established, since : "objectivity can be reached to a certain extent (in tests of a global character) by equalising test procedures and criteria for marking" (9). The criteria selected were established as a function of our aims. They were as follows :

i) Fluency : (tentatively defined as the ability to give proof of sustained oral production implying a certain communicative competence, as well as the unstilted, spontaneous use of English "conversational lubricants" (10).)

ii) Accuracy : (structural and lexical accuracy.)

iii) Relevance : (the ability to talk meaningfully about the topic, to stick to the point, internal consistency, appropriateness, effectiveness.)

iv) Intelligibility : (the overall ability to make oneself understood by the correct use of stress and intonation features.)

v) Pronunciation : (individual pronunciation inadequacies not necessarily of a nature to impair intelligibility.)

These were the most salient criteria reflecting a hierarchy of importance. Further criteria were the following, justified by the nature of the course-orientated topics offered for commentary :

vi) Variety of structure

vii) Variety of lexis.

It might be objected at this stage that in spite of having broken down criteria into reconizable fields one is still confronted with a subjective interpretation within each field. Nevertheless it is felt that some progress has been made on the road to reducing subjectivity to a limited number of well defined areas enabling greater concentration on each subcomponent and therefore less likelihood of oversight (11). The degree of subjectivity is limited by the degree of equalisation

---

9) Gorosch (1971, p. 5).

10) Abercrombie (1963, p. 57).

11) During the trial periods an attempt at transcribing tapes was made with a view to calculating errors as a function of the words spoken, using the formula suggested by Dodson (1963). However this formula only worked if the group was extremely homogeneous. Moreover, attempts at classification of errors in order to weight them made the formula extremely difficult to handle and turned out to be no more objective than the alternative method of scoring adopted. Added to this was the length of the operation which partly invalidated the purpose of replacing interviews.

of criteria and procedures used in evaluation.

As with the Pimsleur experiment (12), scorers are instructed to practice on at least ten recordings in order to stabilize their judgement before beginning actual marking and new markers have their materials checked by an experienced colleaugue before being allowed to score alone. All scores are requested to select a cross-section of weak, average and strong candidates from the range of examination tapes they have been entrusted with for them to be remarked independently by a colleague not aware of the scores already attributed to the tapes. Remarkable consistency has been noted between the different markers so that the operation incidentally tested interjudge reliability. Storage of cross-sections of tapes allows for cross-comparison from year to year.

f) Problems

Several methodological and technical problems need to be taken into account for the successful administration of the above test.

Time allowed for students to speak on a particular topic must be carefully selected in accordance with group levels. Inevitably, an exceptionally strong student will occasionally feel frustration at not being allowed to develop his ideas at greater length. This has no bearing on the reliability or validity of the test. What is more important is that the weaker student manages to fill up the time available without having recourse to repetition.

The impersonal nature of the monologue, as opposed to the dialogue situation implicit in the interview may give rise to certain objections, but these can be answered by the fact that the "Test of Oral Fluency" described here is not an isolated one but is part of a battery. The ability to communicate in dialogue is tested elsewhere, this test acting as a supplement to measure a different aspect of linguistic expression. Here the emphasis is laid on the ability to produce connected discourse but a safeguard against the test becoming an exercise in haranguing or depending too much on imagination is built in with the time-limit imposed.

Great importance must be given to the selection of suitable topics since they must be of a nature likely to elicit a continuous flow of speech. Selection will,

---

12) Pimsleur (1966, p. 197).

of course, depend on the nature of the courses given as will success on the amount of spontaneous speech encouraged throughout teaching periods.

Conclusions.

It is quite clear from the number of restrictions alluded to throughout this paper that the test of oral fluency described is not a perfect solution in the quest for objective standardised tests of productive speech. Much research has still to be undertaken before any such panacea will be found. Important problems such as the difference between mistakes and errors has to be resolved, where errors represent the systematic errors of the learner that represent his "transitional competence", with mistakes designating the unsystematic errors of performance (13). Connected with these is the question of the reliability of error classification.

Nevertheless the test described here is not invalidated by the problems that surround it if one accepts that at the very least it does no less than that which is attainable by the traditional interview situation. Moreover, only a part of the criticisms that can be levelled at interviews can be applied to the above test. Also, it goes further than most of the currently used proficiency tests which measure mechanical rather than communicative skills (14).

On the positive side the experiment has highlighted and classified some of the elements involved in interview testing and has then tried to reduce these features into more controlable items. The testing technique applied has standardised the extralinguistic elements to a far greater extent than in the past. The marking criteria used have brought about "reduced subjectivity" by introducing a "scoring method which maximises agreement among different judges" (15) while at the same time introducing optimum marking conditions. Ease of application, gain in time, and above all, the test's reliability, justify continuing with the experiment. The test also appears to satisfy criteria of validity which future experimentation should reveal.

Work is still in progress on attempts to further standardise marking and thereby narrow the discrepancy ratio even more by applying a rigorous system

---

13) Pit Corder (1967, p. 167).

14) Jakobovits (1971, p. 50).

15) Pimsleur (1966, p. 195).

of + and - symbols to the scoring criteria employed in the hope of eventually converting these symbols into a clear-cut pass/fail or points scale.

The above experiment has been a group project undertaken in the English department of the Institut de Phonétique. It owes its success to the collaboration of E.J. Lee, S.L. Kenny-Levick, R. Most, L. Macilwaine, C.T. Marks and A.C. Woodcock.

October 1972

Dr. H. Baetens Beardsmore

BIBLIOGRAPHY

Abercrombie, D. : Problems and Principles in Language Study, London, Longmans, 1963, 2nd. edition.

ATLA Bulletin : Number 1(9), Jan-March, 1972, Stockholm.

Baetens Beardsmore, H. & Renkin, A. : "A Test of Spoken English", in IRAL, IX/I, 1971, pp. 1-11.

Campbell, R. & Wales, R. : "The Study of Language Acquisition", in New Horizons in Linguistics, ed. J. Lyons, Harmondsworth, Pelican, 1970, pp. 242-260.

Dodson, J. : Oral Examinations, Pamphlet n° 12, Aberystwyth, Dept. of Education, University of Wales, 1963.

Gorosch, M. : "The Linguistic Content, Means of Evaluation and their Interaction in the Teaching and Learning of Modern Languages in Adult Education", Council of Europe Symposium, Rüschlikon, 3-7 May, 1972.

Grayshon, M. : The Examination of Spoken English, Nottingham, University of Nottingham, Institute of Education, 1968.

Jakobovits, L. : Foreign Language Learning, Rowley, Newbury House Publishers, 1971, 2nd. edition.

Macnamara, J. : "How can one measure the extent of a person's bilingual proficiency?" in Kelly, (ed) The Description and Measurement of Bilingualism, University of Toronto Press, 1969.

Pimsleur, P. : "Testing Foreign Language Learning", in Valdman, (ed) Trends in Language Teaching, New York, McGraw-Hill, 1966.

Pit Corder, S.: "The Significance of Learners' Errors", in IRAL, V/4, 1967, pp. 161-170.

Spencer, J. : "Special Languages" : teaching rules or learning roles? Communication read at the ASLA/AIMAV Seminar on Modern Language Teaching to Adults, Stockholm, 27-30 April, 1972.